Minimum-cost Cloud Storage Service Across Multiple Cloud Providers

Guoxin Liu and Haiying Shen

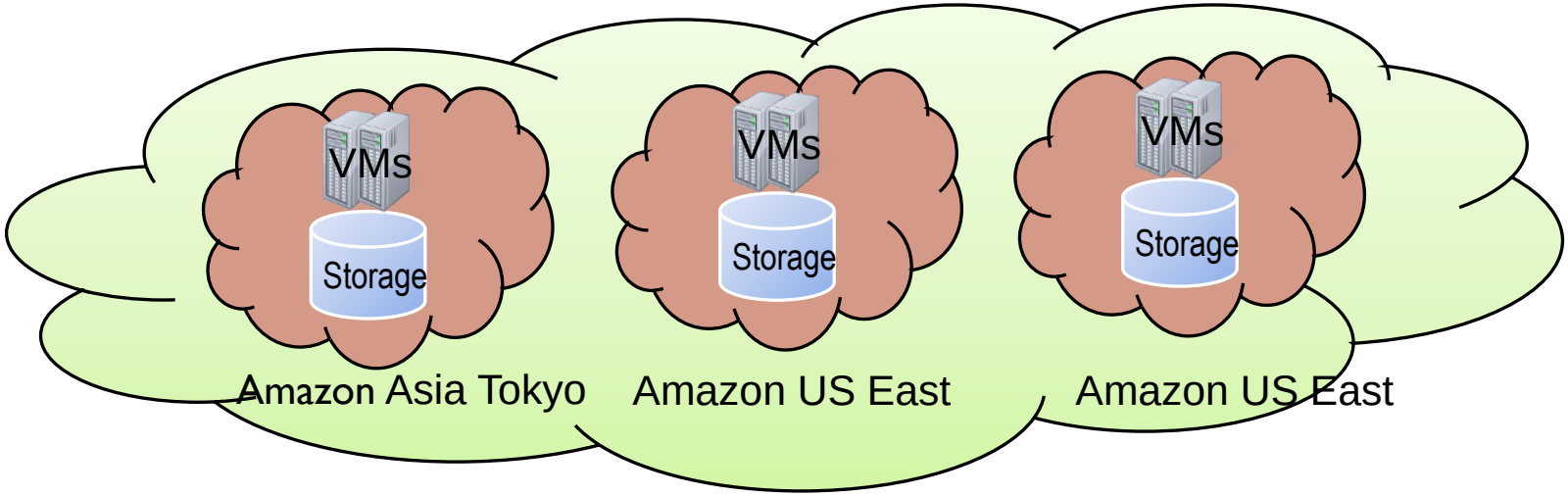
Department of Electrical and Computer Engineering, Clemson University, Clemson, USA

Outline

- Introduction
- Related Work
- Data Storage Service across Cloud Service Providers (CSPs)
- Evaluation
- Conclusion

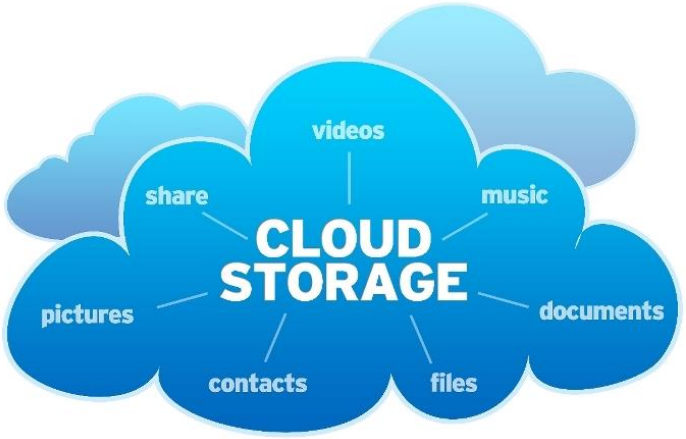
Worldwide Cloud Storage

- Cloud storage
  - Vendors
    - Amazon Dynamo; Microsoft Azure; Google Cloud Storage
  - Benefits
    - Save capital investment
      - World wide distributed datacenters
    - Save management and maintaining cost
    - Pay for usage
      - 1) Pay-as-you-go, 2)Reservation



Worldwide Cloud Storage

- Cloud storage
  - New solution for web application storage*
    - OSNs: Google+
    - Portals: BestBuy
    - File sharing: Dropbox
  - Using a single CSP
    - Vendor lock-in
    - Service latency (SLO)
    - Payment cost
  - Cloud storage across multiple CSPs
    - How to choose vendors
      - Amazon Dynamo
      - Microsoft Azure
      - Google Cloud Storage



* H. Stevens and C. Pettey. Gartner Says Cloud Computing Will Be As Infuential As E-Business. Gartner Newsroom, Online Ed., 2008.

Outline

- Introduction
- Related Work
- Data Storage Service across Cloud Service Providers (CSPs)
- Evaluation
- Conclusion

Related Work

- Deploy on multiple clouds
  - Transparently spread the storage workload over many cloud storage providers
    - Data availability: RACS [1], DepSky [2]
    - Data retrieval latency: COPS [3], Volley [4]
- Minimize Cloud storage cost
  - Adaptively assign data with different sizes to different storage services to minimize the cost for storage
    - Minimize resource utilization of a cluster [5]
    - Minimize cost across CSPs [6]

[1] A. Hussam, P. Lonnie, and W. Hakim. RACS: A Case for Cloud Storage Diversity. In Proc. of SoCC, 2010.
[2] A. N. Bessani, et al. DepSky: Dependable and Secure Storage in a Cloud-of-Clouds. TOS, 2013.
[3] W. Lloyd, et al. Dont Settle for Eventual: Scalable Causal Consistency for Wide-Area Storage with COPS. 2011.
[4] S. Agarwal, et al. Volley: Automated data placement for geo-distributed cloud services. NSDI, 2010.
[5] H. V. Madhyastha, et al. SCC: Cluster Storage Provisioning Informed by Application Characteristics and SLAs.
[6] Z. Wu, et al. SPANStore: Cost-Effective Geo-Replicated Storage Spanning Multiple Cloud Services. SOSP, 2013.

Related Work (cont.)

- SLO guarantee in datacenters
  - Achieve high throughput and guarantee SLO
    - Caching and scheduling [7], [8]
- We propose geo-distributed cloud storage system for Data storage and request Allocation and resource Reservation across multiple CSPs (DAR)
  - Minimize cost considering both pay-as-you-go and reservation pricing models
  - Guarantee SLO considering service latency, data availability and datacenter capacity

[7] C. Hong, M. Caesar, and P. B. Godfrey. Finishing Flows Quickly with Preemptive Scheduling. In Proc. of SIGCOMM, 2012.
[8] B. Vamanan, J. Hasan, and T. N. Vijaykumar. Deadline-Aware Datacenter TCP (D2TCP). In Proc. of SIGCOMM, 2012.

Outline

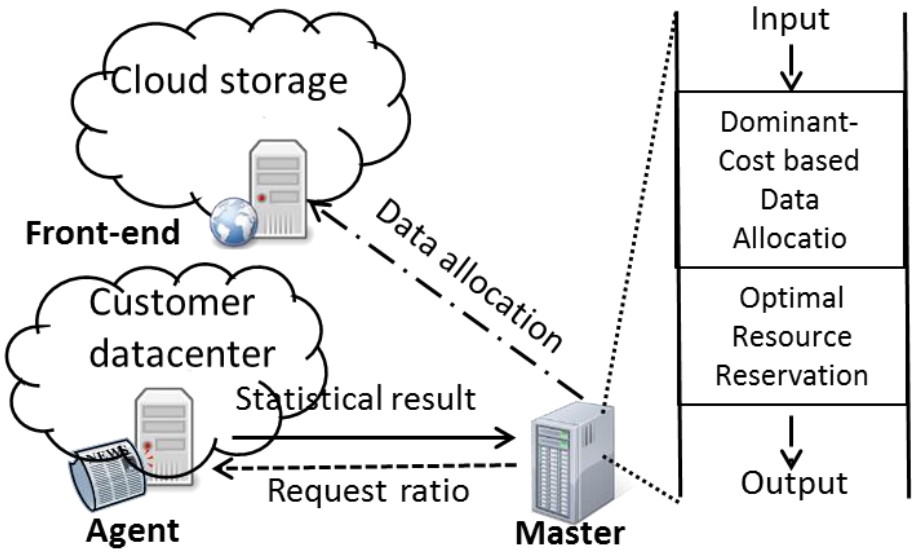
- Introduction
- Related Work
- Data Storage Service across Cloud Service Providers (CSPs)
- Evaluation
- Conclusion

Data Storage Service across CSPs

- Objective: Cost minimization
  - Include: Get, Put, Storage, and Transfer
- Constraints:
  - Service latency
    - Allow no more than % of Gets/Puts beyond required deadline
  - Data availability
    - Maintain a pre-defined number of replicas for each data item
      - Amazon S3 applies 3-replicas policy within a region
  - Datacenter capacity
    - Cumulative Get/Put rate of all data items ≤ Get/put capacity of a datacenter
- Integer Programming
  - NP-hard problem

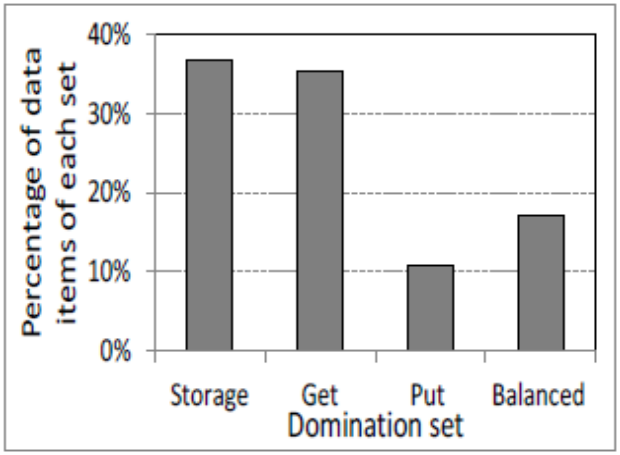
Data Storage Service across CSPs

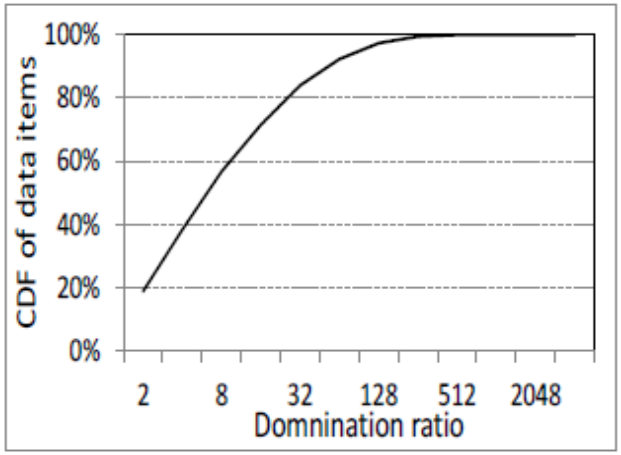
- DAR: Data storage and request Allocation and resource Reservation across multiple CSPs
  - System infrastructure
    - Master and multiple agents
  - A heuristic solution
    - A dominant-cost based data allocation algorithm
    - An optimal resource reservation algorithm



Data Allocation and Reservation

- Dominant cost
  - Cost >> sum of all other costs
  - Storage/Get/Put dominant set
- Trace data analysis
  - Existence & high domination ratio





Data Allocation and Reservation

- Dominant-Cost based Data Allocation
  - Data Allocation for each data item
    - Filter datacenters according to SLO
      - Service latency
      - Data availability
      - Data center capacity
    - Find dominant cost
    - Allocate to datacenters with minimum pay-as-you-go price for dominant cost
- Optimal Resource Reservation
  - Maximize saving in pay-as-you-go cost
  - Avoid over-reservation
  - Proof of the existence of optimal reservation
    - Binary search to find the optimal reservation

Outline

- Introduction
- Related Work
- Data Storage Service across Cloud Service Providers (CSPs)
- Evaluation
- Conclusion

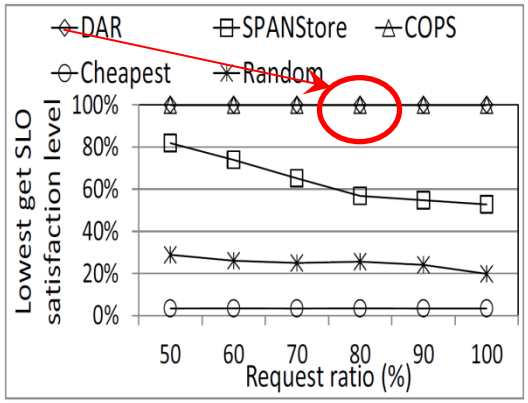
Evaluation of DAR

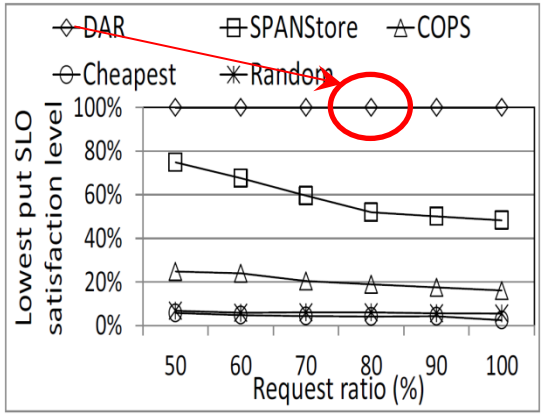
- Simulated CSPs: 25 regions
  - Amazon S3, Microsoft Azure, and Google cloud storage
- Simulated customers
  - 50 Cloud customers
- Real deployment
  - One customer: Amazon EC2 US East & West regions
- Comparison
  - COPS [3]: shortest latency
  - SPANStore [6]: latency guaranteed and unit cost minimization
  - Cheapest: unit cost minimization
  - Random: random CSP selection

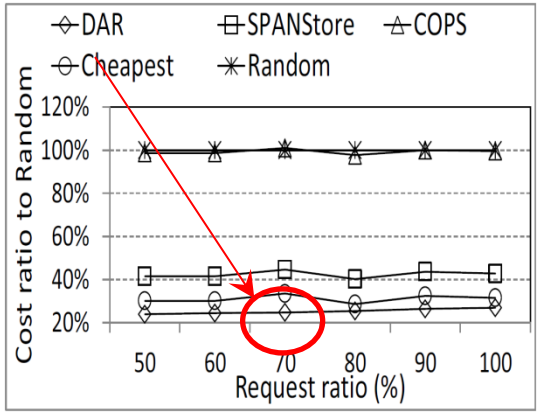
[3] W. Lloyd, et al. Dont Settle for Eventual: Scalable Causal Consistency for Wide-Area Storage with COPS. 2011.
[6] Z. Wu, et al. SPANStore: Cost-Effective Geo-Replicated Storage Spanning Multiple Cloud Services. SOSP, 2013.

Evaluation of DAR

- Due to capacity and latency awareness
  - DAR supplies get-SLA and put-SLA guaranteed service
- Due to the comprehensive cost policy awareness
  - DAR generates the least payment cost to CSPs







Outline

- Introduction
- Related Work
- Data Storage Service across Cloud Service Providers (CSPs)
- Evaluation
- Conclusion

Conclusion

- Data storage and request Allocation and resource Reservation across multiple CSPs (DAR)
  - Algorithms
    - Dominant-cost based data allocation
    - Optimal resource reservation
  - Effectiveness
    - Minimize the payment cost and achieve the SLO of each customer
- Future wok
    - Data request dynamically distribution for congestion control

Thank you!

Questions & Comments?

If you have any questions, please contact:

Haiying (Helen) Shen, Associate professor

shenh@clemson.edu

Clemson University